SCIENTIFIC REPORTS

OPEN

Genomic assemblies of newly sequenced *Trypanosoma cruzi* strains reveal new genomic expansion and greater complexity

Received: 27 April 2018
Accepted: 13 September 2018
Published online: 02 October 2018

Francisco Callejas-Hernández[1], Alberto Rastrojo[1], Cristina Poveda[1], Núria Gironès[1,2] & Manuel Fresno[1,2]

Chagas disease is a complex illness caused by the protozoan *Trypanosoma cruzi* displaying highly diverse clinical outcomes. In this sense, the genome sequence elucidation and comparison between strains may lead to disease understanding. Here, two new *T. cruzi* strains, have been sequenced, Y using Illumina and Bug2148 using PacBio, assembled, analyzed and compared with the *T. cruzi* annotated genomes available to date. The assembly stats from the new sequences show effective improvement of *T. cruzi* genome over the actual ones. Such as, the largest contig assembled (1.3 Mb in Bug2148) in *de novo* attempts and the highest mean assembly coverage (71X for Y). Our analysis reveals a new genomic expansion and greater complexity for those multi-copy gene families related to infection process and disease development, such as Trans-sialidases, Mucins and Mucin Associated Surface Proteins, among others. On one side, we demonstrate that multi-copy gene families are located near telomeric regions of the "chromosome-like" 1.3 Mb contig assembled of Bug2148, where they likely suffer high evolutive pressure. On the other hand, we identified several strain-specific single copy genes that might help to understand the differences in infectivity and physiology among strains. In summary, our results indicate that *T. cruzi* has a complex genomic architecture that may have promoted its evolution.

Trypanosoma cruzi is a highly polymorphic parasite that belongs to the order Kinetoplastidae, being the causative agent of Chagas disease or American Trypanosomiasis, a chronic illness and one of the most neglected tropical diseases[1,2]. Chagas disease is endemic in Latin America but due to migration of infected people, this disease has been extended to non-endemic areas, such as the European Union and the United States, making of Chagas disease a serious public health problem[3–5]. Currently, 7 to 10 million people are thought to be infected and 10,000 to 14,000 deaths per year are caused by this disease[2].

Chagas disease occurs in two phases: acute and chronic, that can be asymptomatic for decades and evolve towards the most aggressive symptomatic phases characterized by cardiac and/or digestive clinical forms[1,2]. This broad spectrum of clinical manifestations had been related to parasite and host genetic variability[6,7]. Thus, there is a significant variation in the size, structure and ploidy among the different parasite strains, although *T. cruzi* is usually described as diploid[8–13]. Genes in trypanosomatids do not present promoter regions to regulate gene expression. Contrarily, gene duplication and mRNA stability, among others, are likely the possible mechanisms to control gene expression[14], as it has been well documented in other Kinetoplastidae as *Leishmania ssp*. and *T. brucei*. This phenomenon explains the selection of highly repetitive (or low complexity) genomic content and variations through evolution[9,15–18], likely originated by the accumulation of transposable elements, high prevalence of tandemly repetitive sequences (involved in gene divergence) and repetitive short sequences in chromatin remodeling[19,20]. Thus, the relationship between short repeats, tandem repeats and genome size has been of interest because some repeated sequences are significantly expanded in length during evolution[20–22]. In addition, the genomic repetitiveness in other species has been associated with natural selection, where low complexity plays a critical role in the control and/or mediation of gene expression and variability[23,24]. In the 19th century, Haldane

[1]Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Científicas, Universidad Autónoma de Madrid, Cantoblanco, Madrid, Spain. [2]Instituto Sanitario de Investigación Princesa, Madrid, Spain. Correspondence and requests for materials should be addressed to N.G. (email: ngirones@cbm.csic.es) or M.F. (email: mfresno@cbm.csic.es)

Strain	DTU	Size (MB)	GC (%)	Contig N50 *	Max lenght (PB)	Min lenght (BP)	Mean contigs (BP)	Contigs	Scaff level*	Coverage (%)	Sequencing method
Dm28c	I	27.34	53.33	78,389	462,134	2,004	22,601	1,210	Scaff*	63	454*
Jrcl4	I	41.48	52.45	83,591	828,981	200	2,709	15,312	Scaff	69	454
Sylvio X10	I	38.58	52.91	2,307	72,500	202	1,428	27,019	Contigs*	30	454 + illumina
Y	II	39.34	51.43	11,782	304,87	500	3,885	10,127	Contigs	71	Illumina
Esmeraldo	II	38.08	52.09	66,229	483,664	200	2,409	15,803	Scaff	60	454
Bug2148	V	55.22	51.63	196,760	1,305,792	585	59,129	934	Contigs	68	PacBio*
Tula	VI	83.51	50.63	7,772	242,476	200	1,826	45,711	Scaff	50	454
BNEL	VI	32.52	43.94	870,934	2,371,736	77,958	793,392	41	Chrom*	14	454 + BAC*
BEL	VI	32.52	40.35	870,934	2,371,736	77,958	793,391	41	Chrom	14	454 + BAC
B7	VI	38.65	52.87	25,781	335,615	235	2,302	16,783	Contigs	30	454 + illumina

Table 1. Summary of *T. cruzi* assembled genomes. *Contig N50: is a statistic median such that the 50% of the entire assembly is contained in contigs equal to or larger than this value. *Scaff level: Scaffolded level assembly. *Chrom: Chromosome level assembly. *454: Roche 454. *PacBio: Pacific Biosciences. *WGS: Whole Genome Sequecing. *BAC: Bacterial Artificial Chromosome. Previous *T. cruzi* genomes available at https://www.ncbi.nlm.nih.gov/genome/genomes/25? Newly sequenced genomes are underlined.

and Stadler suggested that gene duplication and divergence might be favorable because of the possibility to produce genes without any disadvantage to the organism, in the sense that multi-copy genes would be less susceptible to damaging mutations demonstrating that polyploid unicellular organisms were less susceptible than their diploid congeners to irradiation[25].

In 2009 a classification based on the genetic structure (including genomic and mitochondrial kinetoplastid DNA) was proposed that describes the existence of six separated clusters or discrete typing units (DTUs) of *T. cruzi* isolates, or strains, named from TcI to TcVI, where TcV and TcVI have a hybrid evolutionary origin, with TcII and TcIII as putative parents[26]. However, this classification has been questioned after a systematic phylogenetic tree reconstruction based on most frequent mitochondrial *T. cruzi* genes in genome databases, which showed the existence of three significant clusters named as mtTcI, mtTcII and mtTcIII instead 6 DTUs (7 including B7 belonging to TcBat) at least with the actual data and phenotyping techniques[27].

An extensive number of publications about this disease suggest significant phenotypic variation and different behavior both *in vitro* and *in vivo* in terms of pathophysiology, virulence, tropism and immunological responses, which strongly difficult the development of vaccines or new drugs against this disease for which available treatments have limited efficacy and side effects[2,28–30].

Despite these limitations, the publication of several *T. cruzi* genomes has represented an important advance for the understanding of the complexity of this parasite[31]. To date, there are available public genomes from Dm28c, JRCl4 and Sylvio X10 (TcI), Esmeraldo (TcII), Tula, CL Brener Esmeraldo-like "BEL" and non-Esmeraldo-like "BNEL" (TcVI), and B7 (belonging to *T.c marinkellei*) strains, but not from TcIII, TcIV and TcV strains. However, whole genomes or whole chromosomes (Mb of length) cannot be read at a glance with any actual technology, thus millions of fragmented sequences must be assembled in longer fragments named contigs, and when possible, in scaffolds as result of the in tandem union of contigs. Besides, some important questions have raised from genome organization that are still unresolved, such as: (1) the reason and consequences of different genome sizes between strains, (2) karyotype polymorphism even between strains belonging to the same DTU, (3) the presence of tandem repeats and low complexity regions along the whole genome and (4) its relationship with the physio-pathogenesis of the disease. In this study, two *T. cruzi* strains have been sequenced by Next Generation Sequencing (NGS) technologies, assembled with new software programs, analyzed and compared to available genomes. The TcII Y strain was chosen due to their importance in terms of virulence, infectivity and disease development in experimental models[32–34], while Bug2148 belonging to TcV, was chosen because it is more frequently associated to vertical transmission[35].

Results and Discussion

Genome assembly and performance. After quality and length trimming process of the sequences, about 18 million of paired reads for the Y strain were obtained (mean read length: 240 pb), and 757,037 reads for Bug2148 (mean read length: 14 kb), which correspond in both cases to ~10 GB of information and to more than 100x read depth coverage (RDC) for the predicted haploid genomes[36,37]. Contigs from SPAdes assembler for the Y strain (from Illumina) with low coverage (<10X) were removed from the assembly, to reduce chimeric sequences and pseudogenes resulting in 10,127 total assembled contigs (Table 1). In the case of Bug2148 (from PacBio), a total of 934 assembled contigs were obtained with a QV (Quality Value scores) around 99.99% (assembly accuracy) after HGAP's assembly pipeline (Table 1). Genome for the Bug2148 cl1 strain was assembled in a low number of total contigs, lower than almost all available *T cruzi* strains to date. PacBio technology has the capacity of sequence longer reads (until 20 Kb) which in the case of *T. cruzi* genome project may represent a great advantage avoiding complex and/or repetitive regions and facilitating the final assembly. Both new assembled genomes contain around 50% of GC in agreement with previously sequenced *T. cruzi* strains. Based on the total assembled bases, the predicted haploid genome and the total contigs, Bug2148 is probably the most complete haploid *T. cruzi* genome sequenced to date (Table 1).

Strain	Hypothetical Protein (%)
Dm28c	64.26
Sylvio X10	49.50
Y	56.94
Bug2148	53.25
BEL	51.53
BNEL	51.55
B7	50.60

Table 2. Percentage of hypothetical protein content across *T. cruzi* annotated strains.

As an indirect measure of the general genome complexity that may explain the miss-assembly level for the new genomes, the %GC was calculated for each assembled contig (S1 File). In both cases and as it was expected, longest contigs contain around 50% GC composition meanwhile shorter ones possess more variable distribution, indicating assembly breakages on low complexity (more repetitive) genomic areas in agreement with typical *de novo* assemblies (Fig. S1). This pattern of GC distribution is dramatically different for the two new assembled genomes due mainly to the inherent technology differences.

The mean RDC obtained for Y strain assembled genome (71X) is the best mean coverage among all strains actually sequenced and a little higher than for Bug2148 (68X). Unfortunately, complete chromosome reconstruction from short reads produced by Illumina (250 pb for this experiment) is not possible, resulting in very fragmented genomes even in the order of thousands of pieces (Table 1). This is a big problem in complex genomes like Trypanosomatids, and leads to over-, under- or mis-representation of genes or complete chromosomic locations.

In this regard, Arner *et al.*[24] suggested that in *T. cruzi*, the copy number of some conserved genes can be used as misassembly control. For example, monoglyceride lipase gene was predicted to have around 50 copies in several different strains from different DTUs, but despite that, it has been annotated only once for CL Brener and Sylvio X10 genomes (and not found on remaining annotated strains). However, we found 4 copies in the Y strain and 30 in the Bug2148 sequence, which may also confirm the high level of assembly performance in our analysis (Fig. S2).

Gene prediction and functional analysis. Contigs from both new assemblies shorter than 500 pb were removed from the total assembly in addition to RDC filter (10X). Next step was predicting genes across these filtered sequences and predict their theoretical function. A total of 33,306 and 20,058 complete ORFs were obtained for Bug2148 and Y, respectively, being the last figure similar to previous *T. cruzi* gene predictions (belonging to DTUs I and II). However, the number of ORF in Bug2148 was much higher.

Functional annotation of predicted ORFs was performed following two different approaches, first, gene families were defined by self-Blastp and MCL clustering algorithm and second, function was defined by the best reciprocal BLAST against the complete database of protozoan annotated proteins. The minimum e-value was set to 1e-5 and identity $\geq$50%, genes without blast hit were annotated as "Hypothetical protein". After gene clustering process, we obtained a total of 10,549 and 10,674 clusters for Bug2148 and Y, respectively, which correspond to the total and theoretical genes families. For about ~50% of the total predicted genes it was not possible to find a theoretical function, as it is expected for Trypanosomatids (Table 2).

Next step was to perform a genetic diversity analysis for all *T. cruzi* genomes actually sequenced and annotated (available without restrictions at Tritryp data base in 10/06/2017: TcI (Sylvio X10, Dm28c), TcII (**new* Y), TcV (**new* Bug2148 cl1), TcVI (CL Brener BEL and BNEL) and B7. For the new assembled strains ORFs with similarity to known functions were collapsed into families, and protein families for the remaining strains were extracted from their annotated proteins. Results for the 15 more abundant families, which also constitute about 25% of the total genetic content including hypothetical proteins (Y and Bug2148 24.38% and 26.06%, respectively), are shown in Fig. 1A.

In other words, from the total predicted ORFs (33,306 and 20,058 for Bug2148 and Y, respectively) we have identified the total gene families or gene types (10,549 and 10,674 clusters for Bug2148 and Y, respectively) based on its theoretical function. About 50% of the total gene families correspond to hypothetical proteins (also hypothetical conserved) as it was expected. For the remaining 50% of the total genes we have identified 1,589 different functions where the 15 more abundant represent the 25% (approximately) of the total assembled genome (complete data in S2 File), which is lower than previous suggestions of 50%.

Gene clustering for Y strain (Fig. 1B) and Bug2148 (Fig. 1C) highlights the differences between the two new genomes. We confirmed *in silico* the level of divergence between the most abundant protein families that, in some cases have caused its sub-classification. There was a certain degree of divergence between strains respect to the top 5 most abundant known family clusters (in blue circles). This will lead to a better understanding of its relationship with disease development and/or the complex parasite life cycle. Interestingly, we have found that hypothetical proteins may constitute an important and unknown genetic structure of *T. cruzi* since these proteins correspond to at least the 50% of the total *T. cruzi* coding genome that grouped in many clusters and they may constitute new functionally important new families of up 700 copies.

The most abundant gene families in *T. cruzi* are these coding surface protein families such as: Trans-sialidase-like (TS), Mucin, Mucin Associated Surface Protein (MASP), gp63, others related to genetic expression like retrotransposon hot spot (RHS) and the Dispersed Gene Family proteins (DGF) which function remains not well understood. The comparison between available sequences both in absolute copy number as well

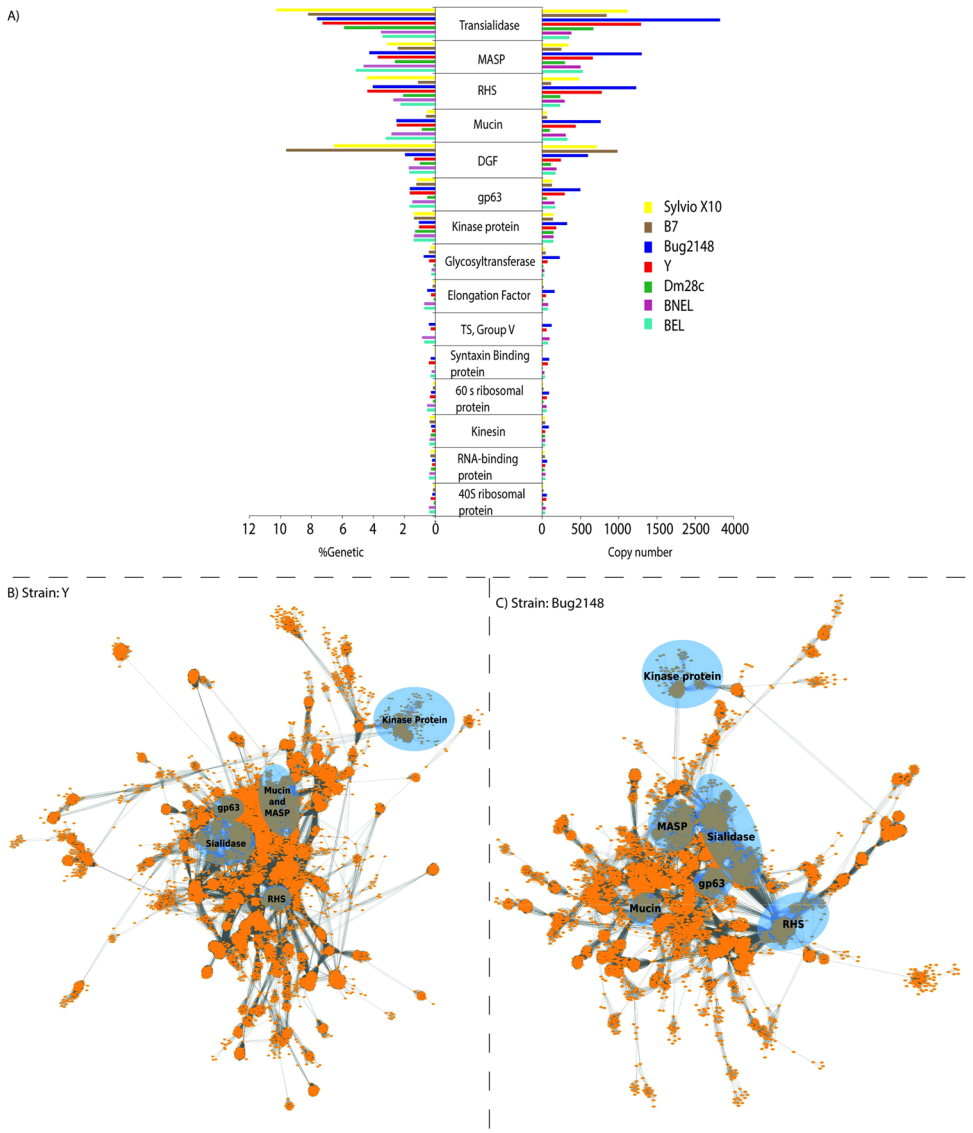

Figure 1. Protein families in *T. cruzi*. (**A**) On the left, the genetic percent that represents the gene copy number found (for the top 15) based on the total annotated genes of Y and Bug2148 and public available genomes. On the right, the copy number for these genes along the entire genome. (**B**) Gene families clustering for the Y strain including hypothetical proteins. (**C**) Gene families clustering for the Bug2148 strain including hypothetical proteins. Most abundant families are remarked by blue circles.

as percentage of that assembled genome revealed interesting differences. The predicted copy number of genes within TS-like family was around 1,400 for Y as it was expected (1,419), and almost the same for Sylvio X10 (1,321), in agreement to previous estimations[38]. In contrast, a higher copy number (2,325) was found in Bug2148 (Fig. 1A, right). However, the percentage that this family represents for each strain from the total genome (coding sequence) is quite different (over to 2% of the total genome, Fig. 1A, left). Some reasons may explain those differences. Although Bug2148 has the highest TS copy number compared to all available strains, the percentage of TS genes respect to the total is very similar to Y and B7 (about the 8%). Thus, differences in copy number are likely due to hybrid origin of Bug2148. As it has been suggested before, the TS family could be under or over-represented due to assembly and technology limitations, including the Y strain, but also, the same tendency may also apply to Mucin, MASP and RHS families. Some genes families encoding the most abundant proteins in *T. cruzi* have been found near repetitive telomeric locations, which may cause collapsed assemblies and therefore very fragmented genomes (shown in Illumina assemblies, Fig. S3)[39]. Moreover, other variables related to the complex kinetoplastid genome such as remarkable karyotype plasticity and aneuploidy are playing decisive roles in sequencing projects[9,28,40].

MASP proteins contain N- and C-terminal domains that encode a putative signal peptide and a GPI-anchor addition site[41]. They constitute about 6% of the parasite haploid genome and are involved in immune evasion[42]. We found here that this figure correlates with the amount predicted for Dm28c (3.5%), TcVI from CL Brener (3.5%), and B7 (3.0%) and the percentage was similar in our Y (3.7%) and Bug2148 (3.9%) strain sequences.

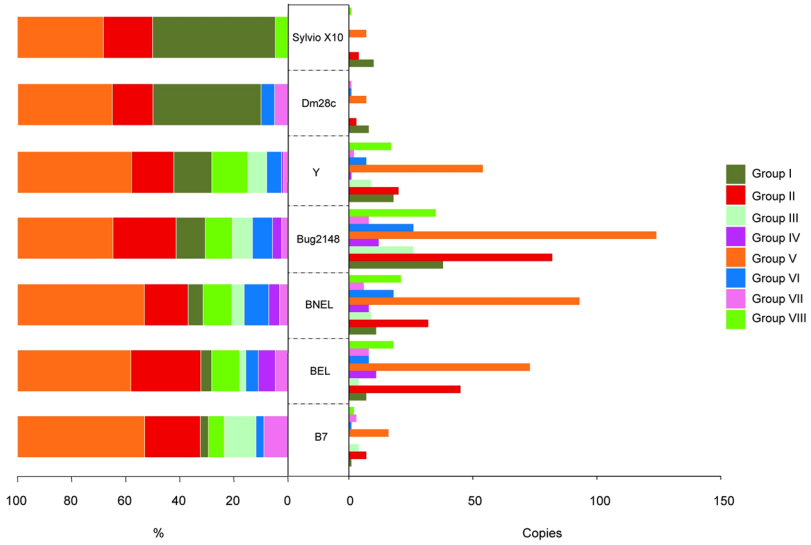

Figure 2. TSs sub-families' distribution (groups I to VIII) in the newly assembled and public genomes. On the left, percent of each TS group based on the total genes that constitutes the complete TS family for each strain. On the right, number of copies for TS groups by strain.

However, for some other multigene families the situation is rather different between strains (Fig. 1A). For example, genes encoding retrotransposon hotspot (RHS) activity, a protein associated to telomeric locations and replicative processes, are as abundant as MASP in Bug2148, Y and Sylvio X10 but much lower in B7, which clearly is not infective in humans. In addition, our analysis showed that mucins are the fourth largest *T. cruzi* gene family (all strains considered), and not the second as it was previously supposed based only on the genetic profile evidence from the CL Brener haplotypes, which were the first annotated genomes. This family is much less represented (2–3 fold) in available Sylvio X10 and Dm28c genomes than in Bug2148, Y and CL Brener. Opposing results were obtained for dispersed gene family proteins (DGF), proteins believed to act similarly to integrins, being the largest family in B7 and also highly abundant in Sylvio X10, 3–4 fold more in percentage than the rest. This family was considered the 5th largest family up to date in *T. cruzi*[42], however, our results showed it may be the largest in the nonpathogenic B7 and the second for Sylvio X10 where constitutes about the 7% of its total coding genome. In addition, this family was suggested to be genetically sub-classified into at least three groups due to its susceptibility to gene recombination promoting gene divergence[43]. However, *in silico* insights showed differential degrees of gene diversity between strains (Fig. S4), more than a general classification clade suitable for all strains. Syntaxin-binding protein family, involved in the regulation of synaptic vesicles docking and fusion processes, is also differentially represented. According to our results, Sylvio X10, Dm28c and B7 have very low percentages of this family, compared with the rest of strains and mainly found in our newly assembled genomes. The reasons and implications of this phenomenon are unknown.

Thus, our results and analysis suggest that to date, the content and diversity of the six more extensive gene families (TSs, MASP, RHS, Mucins, DGF and gp63) is related to a strain-specific genetic profile, the accuracy of its assembled genome, and genomic plasticity variable among strains more than to a general genomic structure pattern. Since these families play multiple roles in virulence, evading vector's and host's defensive mechanisms and are involved in parasite invasion of host cells, the relationship between diversity and expression with strain biological features deserve further and deeper studies.

Trans-sialidase activity: subfamilies. In *T. cruzi*, TS genes comprises a large family of over 1,400 genes sharing the VTVxNVxLYNR motif and TS family members are localized on the membrane surface of metacyclic, blood stream trypomastigotes and intracellular amastigote being the principal gene family involved with host parasite interaction processes[44–46]. Some TS catalyze the transference of sialic acid molecules from host glycoconjugates to acceptor molecules placed on the parasite surface[47], and therefore has been thought to play crucial roles in parasite survival and the establishment of effective infections. In addition, TS is the most polymorphic and complex surface protein family, and also the most abundant genetic family (expressed in the surface) in *T. cruzi*. Genes encoding TS or TS-like genes are actually classified in 8 groups[46], where these groups are defined by specific motifs and show specific activities. Moreover, critical residues necessary for catalytic activity have been identified just in a few genes. Despite this sub-classification and the importance of the family members as virulence factors, most of these subfamilies were not identified in all *T. cruzi* genomes available to date.

Therefore, we analyzed the content of TS subfamilies in all available genomes. Genes from TS group V containing Asp-Box motif (SxDxGxTW) are the most expanded group, both in absolute copy number and % of total TS-like genes, for almost all the strains, except for Dmc28c and Sylvio X10 (of note they belong to the same DTU, TcI) (Fig. 2). This TS group has been associated to antigenic variation, which allows the parasites to adapt to the host environment[46]. TS group II, which does not have enzymatic activity but contains gp85, gp82, gp90 members

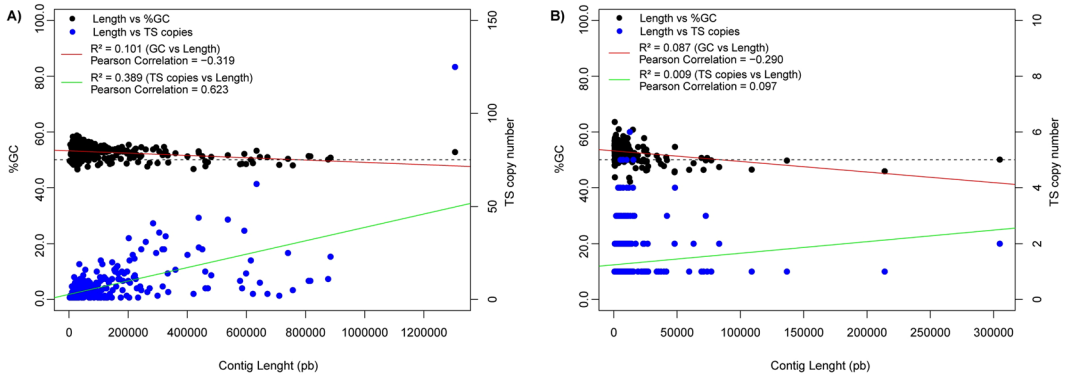

Figure 3. TS-cc length and %GC distribution of the new assembled genomes. Plots of the TS copy number *versus* contig length (blue circles), TS copy number versus %GC (black circles) and Pearson correlations between TS copy numbers and contig length (green line) and between TS copy number and %GC (red dotted line) for Bug2148 (**A**) and Y (**B**) strains.

and other glycoproteins mainly implicated in host cell attachment and invasion, is the second more expanded cluster for Bug2148, BNEL and BEL (124, 93 and 73 copies, respectively).

Interestingly, in the 2 TcI strains Dm28c, Sylvio X10, and in B7 genomes, we could not find TS group IV, although this cannot absolutely discard that they do not exist. It is very important to note that TcI strains including Sylvio X10 are the closest relative to B7 corroborating the high evolutionary relationship of TcBAT with TcI and the emergence of TcBAT genotype in the natural history of *T. cruzi* taxon[48]. TS group V in Y strain showed a low frequency compared to BEL, BNEL and Bug2148, that have the highest percentages and copies.

On the other hand, we found that TS group I, which contains the enzymatically active sialidases TCNA (*T. cruzi* neuraminidase) and SAPA (Shed acute-phase antigen) is not as abundant as predicted, since it was thought to be present in 60–80 copies per haploid genome[49], but according to our analysis of the CL-Brener genome this family is clearly much lower (BNEL 11 copies, BEL 7 copies). In the case of the remaining strains we found that is well represented in the Bug2148 (38 copies), and lower copy number, 18 in the Y strain, 8 and 10 copies for Dm28c and Sylvio X10, respectively. Thus, TS Group I is found in all strains but there are clear differences in copy number and consequently in percentage.

TS Group III includes mostly regulatory proteins that inhibit the alternative and classical complement pathways being FL-160 the best representative[50]. Interestingly, we did not find members of this family in Sylvio X10 and Dmc28c (again the 2 TcI strain). Since FL-160, complement regulatory protein (CRP) family has been also involved in complement system evasion, this may indicate that some strains are more prone than others to evade complement lysis[51]. Therefore, the present gene elucidation and enrichment for this relevant family may contribute to a better understanding of its function.

TS Group IV, which function remains unknown (Tc13 is the representative sequence), was found only present in the CL-Brener genome and in the new Bug2148 strain, being practically absent on other strains that have their genome sequence were more fragmented. Finally, the TS group VII represented by the TS family motif (xTVxx-VxLYNx) was found as one of the smallest group across the annotations. These results confirm that this group is actually not abundant in agreement with previous analysis[9] including the CL-Brener and Bug2148 genomes.

Chromosomal structure in *T. cruzi*. *T. cruzi* has a highly plastic genome, an unusual gene organization and complex mechanisms for gene expression such as polycistronic transcription, RNA editing and trans-splicing[16,52]. Furthermore, due to the lack of mechanisms controlling transcription initiation, subsets of genes must be post-transcriptionally co-regulated in response to extracellular signals. Nowadays, the organization of those genetic subsets remains undefined. More than 50% of the *T. cruzi* genome consists of tandemly repeated sequences and even in diploidy, the parasite has a high variation in chromosome number and even aneuploidy arrangements[9,12,24]. These genome size variations could be related to gene copy number (including pseudogenes and/or non-coding regions), and different reasons had been suggested for this phenomena such as, evasion of the host immune response, the absence of transcription control mechanisms and the complex biological needs through its life cycle in the vector and in the host[53,54].

It has been proposed that some of the most expanded gene families in *T. cruzi* are localized at telomeric and subtelomeric regions that are subject to continuous evolutionary processes. These sites may act as a site for DNA recombination, expansion and generation of new gene variants; this may include DGF, RHS, TS and other acetyltransferases. In particular, TS-like genes have been mainly located near telomeric regions, promoting the generation of new gene variants via non-homologous recombination, as a mechanism by which the parasite evades the host immune response[39]. However, this important evolutionary advantage for the parasite may be causative of *in silico* miss- or over-representation of this complex family through the assembled genomes available to date. This could be due to collapsed assemblies in complex regions, which in some cases correspond to telomeric sequences with specific tandem repeats or other short repetitive genomic motifs. In this sense, we have analyzed the %GC in TS-containing contigs (TS-cc) including pseudogenes, as indirect measure of complexity. Results showed that in both new genomes, short TS-cc have a slightly higher GC content, which negatively correlated to their length (pearson correlation of −0.319 and −0.29 for Bug2148 and Y, respectively) (Fig. 3). This pattern confirms that

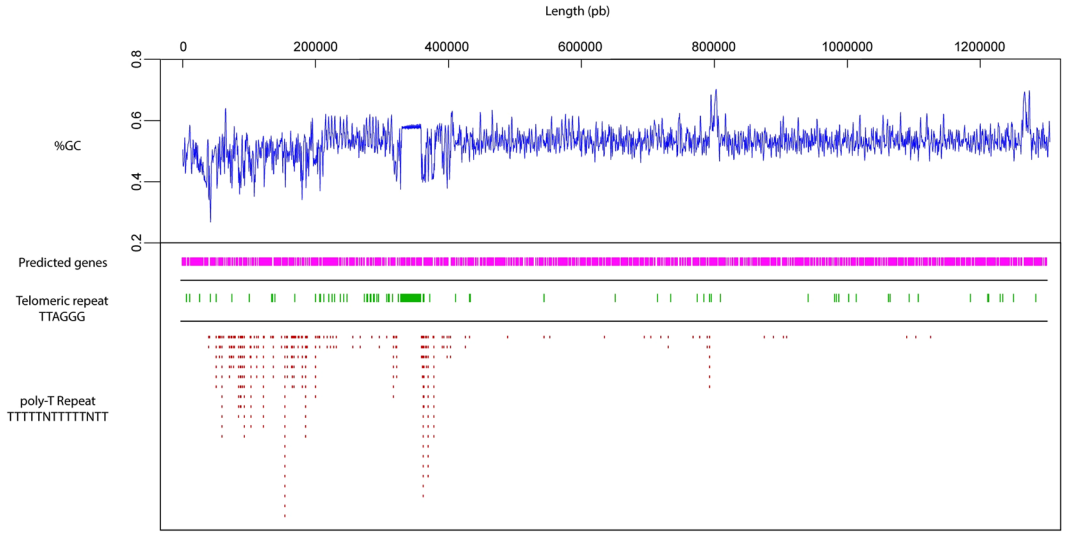

Figure 4. Relation between %GC and telomeric repetitive motifs. %GC variation across largest contig 1.3 Mb ("chromosome-like") from Bug2148 assembly and its relation with the presence of telomeric and poly-T repetitive motifs. TS gene family location at sub-telomeric regions is indicated.

assemblies have collapsed because of highly repetitive sequences mainly GC-enriched, where telomeric motifs may play an important role. However, *T. cruzi* genes (most of them) are packed in tandem arrays or polycistronic transcription units (PTUs) where ORFs may correspond to different functions[55] more than copies in tandem for the same gene. Thus, we have also analyzed the correlation between TS copy number and length in TS-cc, demonstrating that it actually exists in Bug2148 genome specially for contigs about 200–700 Kb long which contains up to 70 TS copies (Pearson correlation of 0.623), but for Y strain there was no significant correlation due to assembly fragmentation, where the longest contig (also identified as TS-cc) barely reached 300 kb.

Once we had strong evidence of TS-cc assembly is affected by complexity, we performed the localization of the specific telomeric repeat (TTAGGG) and poly-T described for *T. cruzi*[56,57] taking as reference the largest contig in Bug2148 (about 1.3 Mb), which correspond to almost the largest entire predicted chromosome in *T. cruzi* (Fig. 4) and also the most important TS-cc (containing 120 copies). This analysis also demonstrates that variations in %GC, at least across this chromosome-like sequence, are directly related to repetitive motifs such as telomers and poly-Ts, in agreement to previous speculations[57], but additionally suggesting that telomeric locations may be 400 kb long. Therefore, this locus has not been previously detected in practically any of the available *T. cruzi* genomes to date, likely due to technology limitations since the De Bruijn Graph algorithms from short sequences used are not the best for reconstruction of long complex sequences.

Interestingly, we found tandem repeats of casein kinase protein (CK) inside the longest tandem chromosomic repeat. This protein is common across all eukaryotic cells, belongs to the multipotential Ser/Thr family and possess the ability of phosphorylate a wide variety of cellular proteins, being one of the most important protein kinases in cell function[58]. However, its specific role in this kinetoplastid remains poorly understood and the reason for this special chromosomal location deserves further and deeper studies.

For the first time, we have described a closer insight to genetic composition in *T. cruzi*, and therefore, we can also confirm that the most important protein families (at least the known ones) were underrepresented in previous genomic sequences due to technical sequencing limitations and/or incomplete assemblies.

Additionally, we have analyzed contigs containing 20 or more telomeric repeats in tandem in the complete genomes and extracted their annotated proteins (Fig. 5). Previous analyses had suggested that about the 9% of TS, 12% of DGF-1, and 19% of RHS were located in this chromosome ends[39]. However, our analyses *in silico* suggests that this figures are underestimated and that chromosome ends include some of the most expanded (with known activity) gene families. Moreover, in agreement to previous analysis we found that about 25% of the second bigger family (MASP) are located in hot-spots in internal chromosomal sites which allow different evolutionary mechanisms as telomeric recombination[41].

Surprisingly, our results show that other protein families, typically considered as less abundant, are located at this complex chromosomic region such as the histones H4, H2A, and the chaperone DNAJ, indicating that these families may be also prone to constant evolution. As it has been described before[59], *T. cruzi* histones present some particularities, mainly at the sequence level, such as high divergence from canonical histones, that in the majority of eukaryotes are highly conserved. In addition, in *T. cruzi*, each histone (canonical and/or variants) is represented by more than one gene. Those coding proteins are described as isoforms, which also may produce unique post-transcriptional modifications marks that are considered essential for chromatin assembly and/or remodelling required in transcription and replication, and in consequence, suggest the existence of essential epigenetic mechanisms in this kinetoplastid[59] that may be also in constant evolution. On the other hand, the complete telomeric chromosomal localization of chaperone DNAJ may explain the low sequence similarity between its gene copies, which is calculated to be around 15–60%.

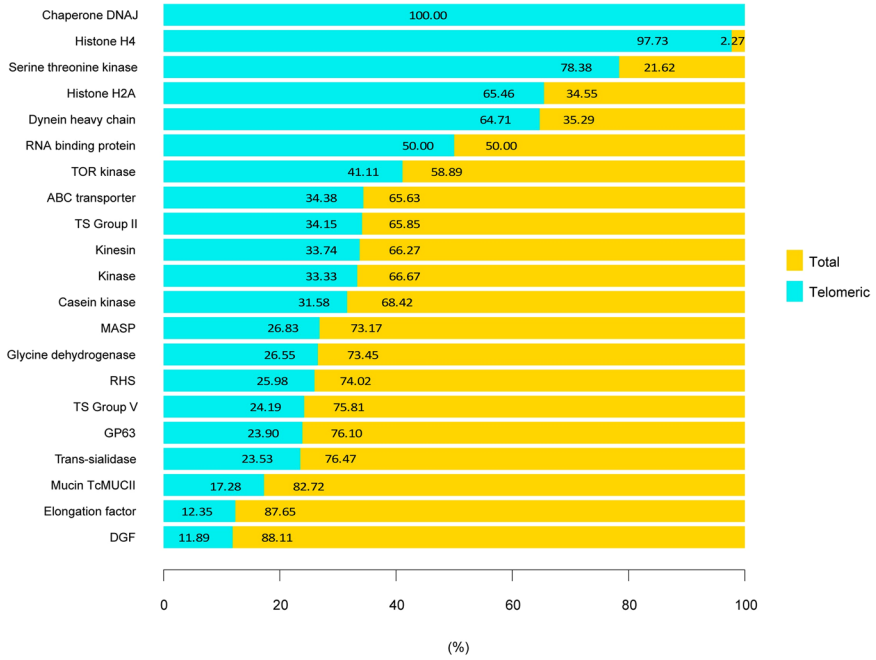

Figure 5. Annotated genes in contigs containing repetitive telomeric motifs. Percentage of annotated genes (by family) near telomeric repeats in Bug2148. Contigs containing telomeric motifs and annotations were not found on Y strain genome.

As we have mentioned before, about 50% of the *T. cruzi* genome is composed of a considerable variety of repeats or complex sequences such as divergent and multi-copy gene families. However, other types of complex non-coding sequences can be found across chromosomes (i.e. genes, SIRE, VIPER, LTR, RLE, SER, etc.), which may also critically affect sequencing projects. Thus, the level of miss-assembly for all those *T. cruzi* genes located within these repeats in the complete genomes available to date is still considerable. In this work, Y strain genome that was sequenced and assembled from short reads (250pb) could have been affected by these obstacles and as a consequence we just found 11 contigs with more than 20 telomeric repeats, where the most abundant proteins located around this motifs (excluding hypothetical proteins) were 3 kinases and one CDS encoding TS activity.

On the other hand, %GC content in some species is also correlated with a number of genomic features potentially relevant functionally such as: gene distribution, transposable elements, methylation rate, and expression levels. GC-rich genes are more efficiently expressed. Genes located in GC-rich regions tend to be also GC-rich in their coding sequences. Also, the average of GC content is higher at silent sites than in neighboring non-coding region, suggesting that high %GC in coding regions could confer some selective advantage[60]. For this reason, we have analyzed contigs containing the highest and lowest %GC distribution among the entire assembly (Fig. 6).

For Y strain genome, a total of 316 contigs with %GC $\leq$35 were found, and 71 of those had at least one CDS (about 22%). However, in the case of Bug2148 23 contigs meeting the mentioned criterion and 17 contains a known predicted function (about 74%). Furthermore, 219 contigs containing %GC $\geq$65 from which 119 have functional annotation were determined for Y, and finally just 2 sequences for Bug2148 (both with annotations corresponding to hypothetical proteins).

In both cases, CDS located in GC-low regions correspond to constitutive functions such as metabolism and gene expression. Interestingly, of those genes at GC-rich locations around half of them correspond to amastigote surface proteins, a highly polymorphic and diverse family.

Predicted single copy genes in Trypanosoma cruzi. Predicted single copy genes (SCG) encoding known protein functions from the two new genomes, were associated to specific biological processes. We found a total of 400 and 183 predicted SCG genes related to specific biological processes for Bug2148 and Y strain, respectively (Fig. S5). This figure is much higher than in previously assembled *T. cruzi* genomes and may have an important impact in the genetic manipulation of this parasite. These differences may be related to assembly performance and/or inherent technology capabilities as we have described before, but due to the potential importance their existence should drive further confirmations. The identification of genetic profiles for each strain, and specifically SCG sequences may define genes that could help to understand differential behaviors trough the parasite life cycle and/or infection cycle. SCG are also more suitable for new genetic manipulation techniques such as CRISPR/Cas9 than multi-copy families constituted by multi-copy sequences. Therefore, they may constitute excellent candidates as drug targets due to their importance in many metabolic pathways.

In addition, the genome sequence from Bug2148 likely allows better resolution and conservation than Y strain, in consequence, the identification of some SCG specific to Y strain not present in Bug2148 would bring interesting outcomes regarding parasite life cycle, virulence or disease evolution of this highly virulent strain.

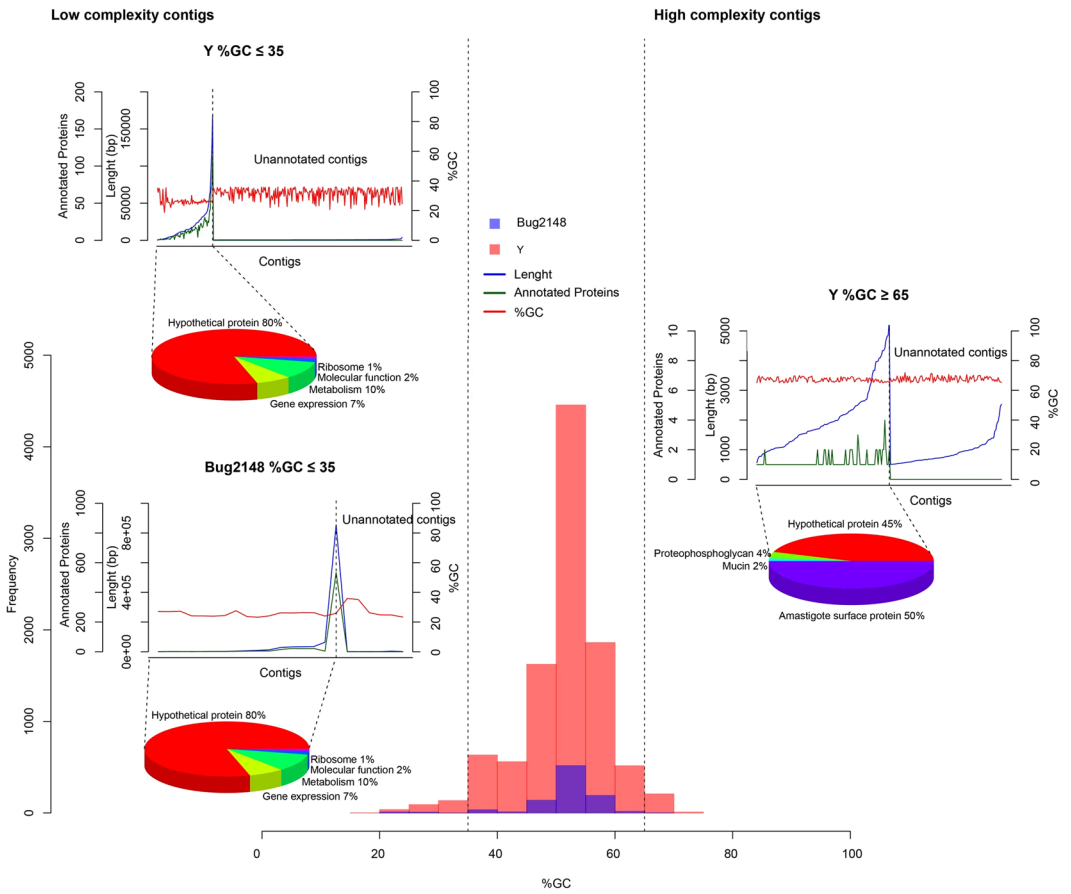

Figure 6. %GC distribution for Y and Bug2148 genomes and genetic composition in the low and high complexity locations. Histogram shows the %GC distribution by contig assembled for the two new genomes, extra graphs shows the genetic annotations on low (lower than 35%) and high (higher than 65%) complexity contigs. (Complete information: File S3).

In this regard, we have identified 2 SCG specific to Y strain: Asparagine-linked glycosylation protein 12 (ALG12) belonging to the carbohydrate derivative biosynthetic process and the molybdenum cofactor biosynthesis protein (Moco) linked to the prosthetic group metabolic process. ALG12 is a prevalent eukaryotic enzyme involved in the co-transcriptional transfer of a pre-assembled tetradecasaccharide from a dolichyl-pyrophosphate donor to the asparagine side chain of maturing proteins in the endoplasmic reticulum (ER)[61]. On the other hand, Moco is a molybdopterin cofactor of xanthine oxidase, DMSO reductase, sulfite oxidase, nitrate reductase, and other oxidases involved in purine metabolism[62]. Notably, xanthine oxidase inhibitors as allopurinol showed trypanostatic effect in infected mice[63].

Conclusions

The Bug2148 genome, based on the *T. cruzi* expected haploid genome size, is the most complete to date, one of the less fragmented assembled by *de novo* reported and the only one belonging to TcV. Thus, it represents a positive contribution for future genomic and transcriptomic (such as RNAseq) analysis.

To date, no thorough comparison of the differences among annotated *T. cruzi* genomes had been performed; important genomic differences between strains and/or DTUs were found in major protein super-families. This reveals a new genomic expansion and complexity and suggests the existence of a species-specific common core genome. The top 15 more expanded families are highly variable among strains. Since these families play also multiple roles in virulence, evading vector's and host's defensive mechanisms, the relationship between diversity and expression with strain biological features opens a new dimension in the studies of *T. cruzi* biology. Those differences may contribute to the understanding of this kinetoplastid, and by extension of one of the most neglected tropical diseases.

We assembled the largest contig (1.3 Mb) that may be considered as the first chromosome assembled (chromosome-like, without scaffolding). As mentioned, genomes (diploid and polyploid) present complex repeat structures including tandem repeats, inverted repeats, imperfect repeats and repeats inserted within repeats[64] that are important but represent a problem for assembly algorithms, giving in most projects very fragmented assemblies of genomes, and therefore a significant loss of information[24,65,66]. In the case of *T. cruzi*, the high frequency of repeats has caused the underrepresentation of the most important gene families implicated in crucial processes, mainly related to vector, host and disease development.

We have provided evidence of gene families and/or structural sequences collapsing assemblies at repetitive telomeric regions of the 1.3 Mb chromosome-like that had not been demonstrated before for *T. cruzi*. These families might suffer more evolutive pressure in those chromosomal locations than others.

Y and Bug2148 new genomes allowed describing many more single copy genes that previously reported (from less than a dozen to several hundreds). For the first time, a closer insight into genetic composition is investigated. Therefore, we believe that our new two *T. cruzi* genomes may positively contribute to move forward in the field.

Methods

Parasite cultures and DNA isolation. Vero cells were grown in RPMI medium supplemented with 5% fetal bovine serum (FBS), 100 UI/mL of antibiotics mixture, 10 µg/mL streptomycin and 2 mM glutamine at 37 °C in an atmosphere of 5% CO_2 until the cells reached 80% confluence (after 4 days). The cell monolayer was subsequently infected with metacyclic trypomastigotes (Bug2148 cl1 and Y, 10 parasites per cell). After 4 days, the supernatant medium was collected, Vero cells and amastigotes were removed by centrifugation at 1000 g by 5 minutes. Trypomastigotes were collected by centrifugation at 1600 g for 10 minutes.

Due to the DNA sample requirements for different sequencing technologies (in length mainly), genomic DNA from Y strain trypomastigotes was isolated using the "High Pure PCR Template Preparation Kit" (Roche) and DNA from Bug2148 was isolated using Phenol-Chloroform method. Both samples were treated with DNAse-free RNAse I (ROCHE) and quantified by absorbance at 260 nm using the Nanodrop ND-1000 (Thermo Scientific). All samples showed an A260/A280 ratio higher than 2.0. In both cases kDNA mitochondrial (shorter than 20 kb) was discarded by agarose gel electrophoresis.

Genome sequencing and data processing. Genome from Bug2148 was sequenced with Pacific Biosciences (PacBio) technologies at the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities[67]. Y strain was sequenced with Illumina MiSeq series by the Genomics facility at the Parque Científico de Madrid (PCM, Madrid, Spain). Integrity from two samples was analyzed in Bioanalizer (Agilent 2100) to confirm DNA fragmentation level larger than 20 Kb for PacBio and 900 bp for Illumina sequencing.

Genome size for DTU's II and V has been estimated to be around 115 and 106 Mb, respectively, where approximately 20% corresponds to kDNA mitochondrial[8,37]. Based on this approach 100x of read depth coverage (RDC) were sequenced for both strains.

No overlapping Paired-end reads of 2 × 300 format and 8–15 kb of read length were obtained from Illumina and PacBio, respectively. Raw reads were subject to quality-filtering using standard processes and analyzed using FASTQC tool[68]. Illumina reads shorter than 50 bp, mean quality lower than 25 (Phred Score based) were removed, and reads longer than 250 bp were trimmed.

Reads shorter than 500 bases, quality lower than 0.8 and polymerase reads shorter than 100 bases were removed from PacBio.

Genome assembly, assembly stats and gene prediction. Trimmed reads from Y strain were assembled using SPADES (v3.9.0)[69], with substrings within a sequence of (k) length (K-mers) for assembly of 21, 33, 55, 77, 99 and 127 K-mers. Bug2146 was assembled using HGAP v3 (Pacific Biosciences, SMRT Analysis Software v2.3.0)[70], seed sequence length and minimal coverage values were set to 6 kb and 15X, respectively.

Statistics from the assemblies were obtained and analyzed by linux command line program Biopieces tool kit[71].

The percentage of GC (%GC) was calculated in each entire contig and in a windowed mode (window = 1000 pb, step = 500 pb). The %GC distribution was obtained with public custom perl scripts[72].

Gene prediction and annotation was performed using Prodigal algorithm (v2.6.3)[73] setting to predict just complete open reading frames (ORFs) by using standard translation eukaryotic table. Gene families were predicted by Markov cluster algorithm (MCL)[74], functional prediction was performed by Best reciprocal BLAST (Basic Local Alignment Search Tool) hit to all proteins available on Tritryp database for *T. cruzi* strains (e-value $\leq$1e-5). Annotations were inspected manually when possible using IGV browser[75]. Single copy genes (SCG) and Monoglyceride lipase were identified and extracted from MCL and Blastp results.

Sialidase genes, functional distribution and telomeric repeats among assembled contigs. Transialidase gene copy quantification, identification of functional subclasses and annotations for complex contigs ($0.35 \leq \%G + C \geq 0.65$) were extracted from MCL and Blastp results, meanwhile telomeric regions were defined by Biopieces (patscan_seq, mismatches not allowed) and IGV Browser (find motif).

Data Availability

Genomes for Bug2148 and Y strain are available from the Genbank database accession numbers NMZN00000000 and NMZO00000000, respectively.

References

1. WHO | Chagas disease (American trypanosomiasis). *WHO* (2017). Available at: http://www.who.int/mediacentre/factsheets/fs340/en/ (Accessed: 4th March 2017).
2. Rassi, A. Jr., Rassi, A. & Marcondes de Rezende, J. American Trypanosomiasis (Chagas Disease). *Infect. Dis. Clin. North Am.* **26**, 275–291 (2012).
3. WHO | Epidemiology. *WHO* (2016). Available at: http://www.who.int/chagas/epidemiology/en/ (Accessed: 11th October 2017).
4. Schmunis, G. A. & Yadon, Z. E. Chagas disease: A Latin American health problem becoming a world health problem. *Acta Trop.* **115**, 14–21 (2010).
5. Bern, C. *et al*. Evaluation and Treatment of Chagas Disease in the United States. *JAMA* **298**, 2171–2181 (2007).

6. Andrade, L. O., Machado, C. R. S., Chiari, E., Pena, S. D. J. & Macedo, A. M. Trypanosoma cruzi: role of host genetic background in the differential tissue distribution of parasite clonal populations. *Exp. Parasitol.* **100**, 269–275 (2002).
7. Manoel-Caetano, F. S. & Silva, A. E. Implications of genetic variability of Trypanosoma cruzi for the pathogenesis of Chagas disease. *Cad Saúde Pública* **23**, 2263–2274 (2007).
8. Souza, R. T. *et al*. Genome size, karyotype polymorphism and chromosomal evolution in Trypanosoma cruzi. *PLoS One* **6**, e23042 (2011).
9. Reis-Cunha, J. L. *et al*. Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct Trypanosoma cruzi strains. *BMC Genomics* **16** (2015).
10. Henriksson, J. *et al*. Chromosomal size variation in Trypanosoma cruzi is mainly progressive and is evolutionarily informative. *Parasitology* **124**, 277–86 (2002).
11. Henriksson, J. *et al*. Chromosomal localization of seven cloned antigen genes provides evidence of diploidy and further demonstration of karyotype variability in Trypanosoma cruzi. *Mol. Biochem. Parasitol.* **42**, 213–23 (1990).
12. Vargas, N., Pedroso, A. & Zingales, B. Chromosomal polymorphism, gene synteny and genome size in T. cruzi I and T. cruzi II groups. *Mol. Biochem. Parasitol.* **138**, 131–41 (2004).
13. Cano, M. I. *et al*. Molecular karyotype of clone CL Brener chosen for the Trypanosoma cruzi Genome Project. *Mol. Biochem. Parasitol.* **71**, 273–278 (1995).
14. Tibayrenc, M. & Telleria, J. *American trypanosomiasis: Chagas disease: one hundred years of research* (Elsevier, 2010).
15. Minning, T. A., Weatherly, D. B., Flibotte, S. & Tarleton, R. L. Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in Trypanosoma cruzi strains revealed by array comparative genomic hybridization. *BMC Genomics* **12**, 139 (2011).
16. De Gaudenzi, J. G., Noe, G., Campo, V. A., Frasch, A. C. & Cassola, A. Gene expression regulation in trypanosomatids. *Essays Biochem* **51**, 31–46 (2011).
17. El-Sayed, N. M. *et al*. Comparative Genomics of Trypanosomatid Parasitic Protozoa. *Science (80-.).* **309**, 404–409 (2004).
18. Castanheira Bartholomeu, D. *et al*. Unveiling the Intracellular Survival Gene Kit of Trypanosomatid Parasites. *PLoS Pathog* **10**, e1004399 (2014).
19. Lorch, Y., Maier-Davis, B. & Kornberg, R. D. Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes Dev.* **28**, 2492–2497 (2014).
20. Hancock, J. M. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. *Genetica* **115**, 93–103 (2002).
21. Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* **24**, 400–402 (2000).
22. Xu, X., Peng, M., Fang, Z. & Xu, X. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24**, 396–399 (2000).
23. Saunders, N. J. *et al*. Repeat-associated phase variable genes in the complete genome sequence of Neisseria meningitidis strain MC58. *Mol. Microbiol.* **37**, 207–215 (2000).
24. Arner, E. *et al*. Database of Trypanosoma cruzi repeated genes: 20,000 additional gene variants. *BMC Genomics* **8**, 391 (2007).
25. Taylor, J. S. & Raes, J. Duplication And Divergence: The Evolution of New Genes and Old Ideas. *Annu. Rev. Genet* **38**, 615–43 (2004).
26. Zingales, B. *et al*. A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. *Mem Inst Oswaldo Cruz Rio Janeiro* **104**, 1051–1054 (2009).
27. Barnabé, C., Mobarec, H. I., Jurado, M. R., Cortez, J. A. & Brenière, S. F. Reconsideration of the seven discrete typing units within the species Trypanosoma cruzi, a new proposal of three reliable mitochondrial clades. *Infect. Genet. Evol.* **39**, 176–186 (2016).
28. Franzén, O. *et al*. Shotgun sequencing analysis of Trypanosoma cruzi i Sylvio X10/1 and comparison with T. cruzi VI CL Brener. *PLoS Negl. Trop. Dis.* **5**, 1–9 (2011).
29. Rodriguez, H. O. *et al*. Trypanosoma cruzi strains cause different myocarditis patterns in infected mice. *Acta Trop.* **139**, 57–66 (2014).
30. Revollo, S. *et al*. Trypanosoma cruzi: Impact of Clonal Evolution of the Parasite on Its Biological and Medical Properties. *Exp. Parasitol.* **89**, 30–39 (1998).
31. Weatherly, D. B., Boehlke, C. & Tarleton, R. L. Chromosome level assembly of the hybrid Trypanosoma cruzi genome. *BMC Genomics* **10** (2009).
32. Calderón, J. *et al*. The Receptor Slamf1 on the Surface of Myeloid Lineage Cells Controls Susceptibility to Infection by Trypanosoma cruzi. *PLoS Pathog* **8**, e1002799 (2012).
33. Cuervo, H. *et al*. Inducible Nitric Oxide Synthase and Arginase Expression in Heart Tissue during Acute Trypanosoma cruzi Infection in Mice: Arginase I Is Expressed in Infiltrating CD68. *J Infect Dis.* **197**, 1772–1782 (2008).
34. Cuervo, H. *et al*. Myeloid-Derived Suppressor Cells Infiltrate Myeloid-Derived Suppressor Cells Infiltrate the Heart in Acute Trypanosoma cruzi Infection. *J. Immunol.* **187**, 2656–2665 (2011).
35. Carlier, Y. & Torrico, F. Congenital infection with Trypanosoma cruzi: from mechanisms of transmission to strategies for diagnosis and control. *Rev. Soc. Bras. Med. Trop.* **36**, 767–771 (2003).
36. Souza, R. T. *et al*. Genome Size, Karyotype Polymorphism and Chromosomal Evolution in Trypanosoma cruzi. https://doi.org/10.1371/journal.pone.0023042 (2011).
37. Lewis, M. D. *et al*. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in Trypanosoma cruzi populations and expose contrasts between natural and experimental hybrids. *Int. J. Parasitol.* **39**, 1305–1317.
38. Oliveira, I. A., Freire-de-Lima, L., Penha, L. L., Dias, W. B. & Todeschini, A. R. Trypanosoma cruzi Trans-Sialidase: Structural Features and Biological Implications. in *Proteins and Proteomics of Leishmania and Trypanosoma* (eds Santos, A. L. S., Branquinha, M. H., d'Avila-Levy, C. M., Kneipp, L. F. & Sodré, C. L.) 181–201 https://doi.org/10.1007/978-94-007-7305-9_8 (Springer Netherlands, 2014).
39. Moraes Barros, R. R. *et al*. Anatomy and evolution of telomeric and subtelomeric regions in the human protozoan parasite Trypanosoma cruzi. *BMC Genomics* **13** (2012).
40. De Bustos, A., Cuadrado, A. & Jouve, N. Sequencing of long stretches of repetitiveDNA. *Nat. Publ. Gr.* **6** (2016).
41. Bartholomeu, D. C. *et al*. Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen Trypanosoma cruzi. *Nucleic Acids Res.* **37**, 3407–3417 (2009).
42. De Pablos, L. M. & Osuna, A. Multigene families in Trypanosoma cruzi and their role in infectivity. *Infect. Immun.* **80**, 2258–2264 (2012).
43. Kawashita, S. Y., Da Silva, C. V., Mortara, R. A., Burleigh, B. A. & Briones, M. R. S. Homology, paralogy and function of DGF-1, a highly dispersed Trypanosoma cruzi specific gene family and its implications for information entropy of its encoded proteins. *Mol. Biochem. Parasitol.* **165**, 19–31 (2009).
44. Frasch, A. C. C. Functional Diversity in the Trans-sialidase and Mucin Families in Trypanosoma cruzi. *Parasitol. Today* **16**, 282–286 (2000).
45. Chiurillo, M. A. *et al*. The diversity and expansion of the trans-sialidase gene family is a common feature in Trypanosoma cruzi clade members. *MEEGID* **37**, 266–274 (2016).
46. Freitas, L. M. *et al*. Genomic Analyses, Gene Expression and Antigenic Profile of the Trans-Sialidase Superfamily of Trypanosoma cruzi Reveal an Undetected Level of Complexity. *PLoS One* **6**, e25914 (2011).

47. Nardy, A. F. F. R., Freire-de-Lima, C. G., Pérez, A. R. & Morrot, A. Role of Trypanosoma cruzi Trans-sialidase on the escape from host immune surveillance. *Front. Microbiol.* **7** (2016).
48. Ramírez, J. D. *et al.* Trypanosome species in neo-tropical bats: Biological, evolutionary and epidemiological implications. *Infect. Genet. Evol.* **22**, 250–256 (2014).
49. Cremona, M. L., Campetella, O., Sánchez, D. O. & Frasch, A. C. Enzymically inactive members of the trans-sialidase family from Trypanosoma cruzi display beta-galactose binding activity. *Glycobiology* **9**, 581–587 (1999).
50. Mathieu-Daudé, F. *et al.* Exploring the FL-160-CRP gene family through sequence variability of the complement regulatory protein (CRP) expressed by the trypomastigote stage of Trypanosoma cruzi. *Infect. Genet. Evol.* **8**, 258–266 (2008).
51. Ramírez-Toloza, G. & Ferreira, A. Trypanosoma cruzi evades the complement system as an efficient strategy to survive in the mammalian host: The specific roles of host/parasite molecules and Trypanosoma cruzi calreticulin. *Front. Microbiol.* **8** (2017).
52. Araújo, P. R. *et al.* Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in Trypanosoma cruzi -A Review. *Mem Inst Oswaldo Cruz* **106**, 257–266 (2011).
53. Najib, M. *et al.* The Genome Sequence of Trypanosoma cruzi, Etiologic Agent of Chagas Disease. *Science (80-.).* **309**, 409–415 (2005).
54. Tomás, A. M. & Kelly, J. M. Stage-regulated expression of cruzipain, the major cysteine protease of Trypanosoma cruzi is independent of the level of RNA. *Mol. Biochem. Parasitol.* **76**, 91–103 (1996).
55. Vazquez, M. P. The Genetics and Genomics of Trypanosoma cruzi. *Int. J. Biomed. Pharm. Sci.* **1**, 1–11 (2007).
56. Kim, D. *et al.* Telomere and subtelomere of Trypanosoma cruzi chromosomes are enriched in (pseudo)genes of retrotransposon hot spot and trans-sialidase-like gene families: the origins of T. cruzi telomeres. *Gene* **14**, 153–161 (2005).
57. Chiurillo, M. A., Isabel Cano, J. F. D. S. & Ramirez., J. L. Organization of telomeric and sub-telomeric regions of chromosomes from the protozoan parasite Trypanosoma cruzi. *Mol. Biochem. Parasitol.* **100**, 173–183 (1999).
58. Turowec, J. P. *et al.* Chapter 23 - Protein Kinase CK2 Is a Constitutively Active Enzyme that Promotes Cell Survival: Strategies to Identify CK2 Subsstrates and Manipulate its Activity in Mammalian Cells. *Methods Enzymol. Const. Act. Recept. Other Proteins, Part A* **484**, 471–493 (2010).
59. Picchi, G. F. A. *et al.* Post-translational Modifications of Trypanosoma cruzi Canonical and Variant Histones. https://doi.org/10.1021/acs.jproteome.6b00655.
60. Kudla, G., Lipinski, L., Caffin, F., Helwak, A. & Zylicz, M. High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *Plos Biol.* **4**, e180 (2006).
61. Eranthie Weerapana and Barbara Imperiali. Asparagine-linked protein glycosylation: from eukaryotic to prokaryotic systems. *Glycobiology* **16**, 91–101 (2006).
62. molybdopterin cofactor biosynthetic process Gene Ontology Term (GO:0032324). Available at: http://www.informatics.jax.org/vocab/gene_ontology/GO:0032324 (Accessed: 23rd November 2017).
63. Avila, J. L. & Avila, A. Trypanosoma cruzi: Allopurinol in the Treatment of Mice with Experimental Acute Chagas Disease. *Exp. Parasitol.* **51**, 204–208 (1981).
64. Miller, J. R., Koren, S., Sutton, G. & Craig, J. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics* **95**, 315–327 (2010).
65. Eichler, E. E. Segmental Duplications: What' s Missing, Misassigned, and Misassembled — and Should We Care? *Genome Res.* 653–656, https://doi.org/10.1101/gr.188901.floor (2001).
66. Salzberg, S. L. & Yorke, J. A. Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320–4321 (2005).
67. Callejas-Hernández, F., Gironès, N. & Fresno, M. Genome Sequence of Trypanosoma cruzi Strain Bug2148. *Genome Announc* **6**, e01497–17 (2018).
68. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ (Accessed: 6th March 2017).
69. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
70. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
71. Biopieces by maasha. Available at: http://maasha.github.io/biopieces/ (Accessed: 6th March 2017).
72. Caballero, J. SeqComplex by jcaballero. Available at: http://caballero.github.io/SeqComplex/ (Accessed: 5th April 2017).
73. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11** (2010).
74. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
75. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29** (2011).

Acknowledgements

This work was supported by the "Consejo Nacional de Ciencia y Tecnología" (CONACYT, México) through the FC-H Ph.D. studentship number 411595 and the "Consejo de Ciencia, Tecnología e Innovación de Hidalgo (CITNOVA, México);; "Ministerio de Economía y competitividad" (SAF2015-63868-R (MINECO/FEDER) to N.G., SAF2016-75988-R (MINECO/FEDER) to M.F.); "Red de Investigación de Centros de Enfermedades Tropicales" (RICET RD12/0018/0004 to M.F.); European Union (HEALTH-FE-2008-22303, ChagasEpiNet to M.F.); Comunidad de Madrid (S-2010/BMD-2332 to M.F.); and Institutional grants from "Fundación Ramón Areces" and "Banco de Santander". The authors would like to thank Maria. A. Chorro and Maria. C. Maza for their excellent technical assistance, to Marcos Parras and Samira Islas for their valuable scientific discussions, Dr. Ricardo Ramos responsible of the "Unidad de Genómica del Parque Científico de Madrid", Dr. Begoña Aguado and Fernando Carrasco responsibles of the "Servicio de Genómica del CBMSO (CSIC/UAM)" for their valuable advice on both sequencing projects.

Author Contributions

C.-H.F. and R.A. performed bioinformatics analysis, C.-H.F. and P.C. prepared cultures and DNA, C.-H.F., G.N. and F.M. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-32877-2.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.